Quantifying the accuracy of approximate diffusions and Markov chains

Jonathan H. Huggins
CSAIL, MIT

James Zou
Stanford University

Abstract

Markov chains and diffusion processes are indispensable tools in machine learning and statistics that are used for inference, sampling, and modeling. With the growth of large-scale datasets, the computational cost associated with simulating these stochastic processes can be considerable, and many algorithms have been proposed to approximate the underlying Markov chain or diffusion. A fundamental question is how the computational savings trade off against the statistical error incurred due to approximations. This paper develops general results that address this question. We bound the Wasserstein distance between the equilibrium distributions of two diffusions as a function of their mixing rates and the deviation in their drifts. We show that this error bound is tight in simple Gaussian settings. Our general result on continuous diffusions can be discretized to provide insights into the computational–statistical trade-off of Markov chains. As an illustration, we apply our framework to derive finite-sample error bounds of approximate unadjusted Langevin dynamics. We characterize computation-constrained settings where, by using fast-to-compute approximate gradients in the Langevin dynamics, we obtain more accurate samples compared to using the exact gradients. Finally, as an additional application of our approach, we quantify the accuracy of approximate zig-zag sampling. Our theoretical analyses are supported by simulation experiments.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

1 Introduction

Markov chains and their continuous-time counterpart, diffusion processes, are ubiquitous in machine learning and statistics, forming a core component of the inference and modeling toolkit. Since faster convergence enables more efficient sampling and inference, a large and fruitful literature has investigated how quickly these stochastic processes converge to equilibrium. However, the tremendous growth of large-scale machine learning datasets – in areas such as social network analysis, vision, natural language processing and bioinformatics – have created new inferential challenges. The large-data setting highlights the need for stochastic processes that are not only accurate (as measured by fast convergence to the target distribution), but also computationally efficient to simulate. These computational considerations have led to substantial research efforts into approximating the underlying stochastic processes with new processes that are more computationally efficient [5, 21, 40].

As an example, consider using Markov chain Monte Carlo (MCMC) to sample from a posterior distribution. In standard algorithms, each step of the Markov chain involves calculating a statistic that depends on all of the observed data (e.g. a likelihood ratio to set the rejection rate in Metropolis-Hastings or a gradient of the log-likelihood as in Langevin dynamics). As data sets grow larger, such calculations increasingly become the computational bottleneck. The need for more scalable sampling algorithms has led to the development of Markov chains which only approximate the desired statistics at each step – for example, by approximating the gradient or sub-sampling the data – and hence are computationally more efficient [4, 5, 13, 21, 26, 27, 33, 40]. The trade-off is that the approximate chain often does not converge to the desired equilibrium distribution, which, in many applications, could be the posterior distribution of some latent parameters given all of the observed data. Therefore, a central question of both theoretical and practical importance is how to quantify the deviation between the equilibrium distribution that the approxi-

mate chain converges to and the desired distribution targeted by the original chain. Moreover, we would like to understand, given a fixed computational budget, how to design approximate chains that generate the most accurate samples.

Our contributions. In this paper, we develop general results to quantify the accuracy of approximate diffusions and Markov chains and apply these results to characterize the computational–statistical trade-off in specific algorithms. Our starting point is continuous-time diffusion processes because these are the objects which are discretized to construct many sampling algorithms, such as the unadjusted and Metropolis-adjusted Langevin algorithms [34] and Hamiltonian Monte Carlo [31]. Given two diffusion processes, we bound the deviation in their equilibrium distributions in terms of the deviation in their drifts and the rate at which the diffusion mixes (Theorem 3.1). Moreover, we show that this bound is tight for certain Gaussian target distributions. These characterizations of diffusions are novel and are likely of more general interest beyond the inferential settings we consider. We apply our general results to derive a finite-sample error bound on a specific unadjusted Langevin dynamics algorithm (Theorem 5.1). Under computational constraint, the relevant trade-off here is between computing the exact log-likelihood gradient for few iterations or computing an approximate gradient for more iterations. We characterize settings where the approximate Langevin dynamics produce more accurate samples from the true posterior. We illustrate our analyses with simulation results. In addition, we apply our approach to quantify the accuracy of approximations to the zig-zag process, a recently-developed non-reversible sampling scheme.

Paper outline. We introduce the basics of diffusion processes and other preliminaries in Section 2. Section 3 discusses the main results on bounding the error between an exact and perturbed diffusion. We describe the main ideas behind our analyses in Section 4; all the detailed proofs are deferred to the Supplementary Material. Section 5 applies the main results to derive finite sample error bounds for unadjusted Langevin dynamics and illustrates the computational–statistical trade-off. Section 6 extends our main results to quantify the accuracy of approximate piecewise deterministic Markov processes, including the zig-zag process. Numerical experiments to complement the theory are provided in Section 7. We conclude with a discussion of how our results connect to the relevant literature and suggest directions for further research.

2 Diffusions and preliminaries

Let $\mathcal{X} = \mathbb{R}^d$ be the parameter space and let π be a probability density over $\mathbb{R}^d$ (e.g. it can be the posterior distribution of some latent parameters given data). A Langevin diffusion is characterized by the stochastic differential equation

$$\mathrm{d}X_t = \nabla \log \pi(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t,$$

where $X_t \in \mathbb{R}^d$ and W_t is a standard Brownian motion. The intuition is that X_t undergoes a biased random walk in which it is more likely to move in directions that increase the density. Under appropriate regularity conditions, as $t \to \infty$, the distribution of X_t converges to π. Thus, simulating the Langevin diffusion provides a powerful framework to sample from the target π. To implement such a simulation, we need to discretize the continuous diffusion into finite-width time steps. For our main results, we focus on analyzing properties of the underlying diffusion processes. This allows us to obtain general results which are independent of any particular discretization scheme.

Beyond Langevin dynamics, more general diffusions can take the form

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{2.1}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ is the drift and is not necessarily the gradient of some log-density.[1] Furthermore, we can analyze other continuous-time Markov processes such as piecewise deterministic Markov processes (PDMPs). For example, Hamiltonian Monte Carlo [31] can be viewed as approximating a PDMP and the zig-zag process is a recently-developed non-reversible PDMP designed for large Bayesian inference (see Section 6).

In many large-data settings, computing the drift $b(X_t)$ in Eq. (2.1) can be expensive; for example, computing $b(X_t) = \nabla \log \pi(X_t)$ requires using all of the data and may involve evaluating a complex function such as a differential equation solver. Many recent algorithms have been proposed where we replace b with an approximation $\tilde{b}$. Such an approximation changes the underlying diffusion process to

$$\mathrm{d}\tilde{X}_t = \tilde{b}(\tilde{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\tilde{W}_t, \tag{2.2}$$

where $\tilde{W}_t$ is a standard Brownian motion. In order to understand the quality of different approximations, we need to quantify how the equilibrium distribution

[1] All of our results can be extended to more general diffusions on a domain $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathrm{d}X_t = b(X_t) + \Sigma\,\mathrm{d}W_t - n_t L(\mathrm{d}t)$, where Σ is the covariance of the Brownian motion, and $n_t L$ captures the reflection forces at the boundary $\partial\mathcal{X}$. To keep the exposition simple, we focus on the simpler diffusion in the main text.

of Eq. (2.1) differs from the equilibrium distribution of Eq. (2.2). We use the standard *Wasserstein metric* to measure this distance.

Definition. The *Wasserstein distance* between distributions π and $\tilde{\pi}$ is

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) = \sup_{\phi \in C_L(\mathbb{R}^d)} |E_\pi[\phi] - E_{\tilde{\pi}}[\phi]|,$$

where $C_L(\mathbb{R}^d)$ is the set of continuous functions $\phi : \mathbb{R}^d \to \mathbb{R}$ with Lipschitz constant $\|\phi\|_L \leq 1$.[2]

The distance between π and $\tilde{\pi}$ should depend on how good the drift approximation is, which can be quantified by $\|b - \tilde{b}\|_2$.[3] It is also natural for the distance to depend on how quickly the original diffusion with drift b mixes, since the faster it mixes, the less time there is for the error to accumulate. Geometric contractivity is a useful property which quantifies fast-mixing diffusions. For each $x \in \mathbb{R}^d$, let $\mu_{x,t}$ denote the law of $X_t \,|\, X_0 = x$.

Assumption 2.A (Geometric contractivity). *There exist constants $C > 0$ and $0 < \rho < 1$ such that for all $x, x' \in \mathbb{R}^d$,*

$$d_{\mathcal{W}}(\mu_{x,t}, \mu_{x',t}) \leq C\|x - x'\|_2 \rho^t.$$

Geometric contractivity holds in many natural settings. Recall that a twice continuously-differentiable function ϕ is *k-strongly concave* if for all $x, x' \in \mathbb{R}^d$

$$(\nabla\phi(x) - \nabla\phi(x')) \cdot (x - x') \leq -k\|x - x'\|_2^2. \quad (2.3)$$

When $b = \nabla \log \pi$ and $\log \pi$ is k-strongly concave, the diffusion is exponentially ergodic with $C = 1$ and $\rho = e^{-k}$ (this can be shown using standard coupling arguments [10]). In fact, exponential contractivity also follows if Eq. (2.3) is satisfied when x and x' are far apart and $\log \pi$ has "bounded convexity" when x and x' are close together [18]. Alternatively, Hairer et al. [23] provides a Lyapunov function-based approach to proving exponential contractivity.

To ensure that the diffusion and the approximate diffusion are well-behaved, we impose some standard regularity properties.

Assumption 2.B (Regularity conditions). *Let π and $\tilde{\pi}$ denote the stationary distributions of the diffusions in Eq. (2.1) and Eq. (2.2), respectively.*

1. *The target density satisfies $\pi \in C^2(\mathbb{R}^d, \mathbb{R})$ and $\int x^2 \pi(\mathrm{d}x) < \infty$. The drift satisfies $b \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ and $\|b\|_L < \infty$.*
2. *The approximate drift satisfies $\tilde{b} \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ and $\|\tilde{b}\|_L < \infty$.*
3. *If a function $\phi \in C(\mathbb{R}^d, \mathbb{R})$ is π-integrable then it is $\tilde{\pi}$-integrable.*

Here $C^k(\mathbb{R}^m, \mathbb{R}^n)$ denotes the set of k-times continuously differentiable functions from $\mathbb{R}^m$ to $\mathbb{R}^n$ and $C(\mathbb{R}^m, \mathbb{R}^n)$ is the set of all Lebesgue-measurable function from $\mathbb{R}^m$ to $\mathbb{R}^n$. The only notable regularity condition is (3). In the Supplementary Material, we discuss how to verify it and why it can safely be treated as a mild technical condition.

3 Main results

We can now state our main result, which quantifies the deviation in the equilibrium distributions of the two diffusions in terms of the mixing rate and the difference between the diffusions' drifts.

Theorem 3.1 (Error induced by approximate drift)**.** *Let π and $\tilde{\pi}$ denote the invariant distributions of the diffusions in Eq. (2.1) and Eq. (2.2), respectively. If the diffusion Eq. (2.1) is exponentially ergodic with parameters C and ρ, the regularity conditions of Assumption 2.B hold, and $\|b - \tilde{b}\|_2 \leq \epsilon$, then*

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) \leq \frac{C\epsilon}{\log(1/\rho)}. \quad (3.1)$$

Remark 3.2 (*Coherency of the error bound*). To check that the error bound of Eq. (3.1) has coherent dependence on its parameters, consider the following thought experiment. Suppose we change the time scale of the diffusion from t to $s = at$ for some $a > 0$. We are simply *speeding up* or *slowing down* the diffusion process depending on whether $a > 1$ or $a < 1$. Changing the time scale does not affect the equilibrium distribution and hence $d_{\mathcal{W}}(\pi, \tilde{\pi})$ remains unchanged. After time s has passed, the exponential contraction is ρ^{at} and hence the effective contraction constant is ρ^a instead of ρ. Moreover, the drift at each location is also scaled by a and hence the drift error is ϵa. The scaling a thus cancels out in the error bound, which is desirable since the error should be independent of how we set the time scale. □

Remark 3.3 (*Tightness of the error bound*). We can choose b and $\tilde{b}$ such that the bound in Eq. (3.1) is an equality, thus showing that, under the assumptions considered, Theorem 3.1 cannot be improved. Let $\pi(x) = \mathcal{N}(x; \mu, \sigma^2 I)$ be the Gaussian density with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\sigma^2 I$ and let $\tilde{\pi}(x) = \mathcal{N}(x; \tilde{\mu}, \sigma^2 I)$. The Wasserstein distance between two Gaussians with the same covariance is the distance between their means, so $d_{\mathcal{W}}(\pi, \tilde{\pi}) = \|\mu - \tilde{\mu}\|_2$. Consider the corresponding diffusions where $b = \nabla \log \pi$

[2] Recall that the Lipschitz constant of function $\phi : \mathbb{R}^d \to \mathbb{R}$ is $\|\phi\|_L \triangleq \sup_{x,y \in \mathbb{R}^d} \frac{\|\phi(x) - \phi(y)\|_2}{\|x - y\|_2}$.

[3] For a function $\phi : \mathbb{R}^n \to \mathbb{R}^m$, define $\|\phi\|_2 \triangleq \sup_{x \in \mathbb{R}^n} \|\phi(x)\|_2$.

and $\tilde{b} = \nabla \log \tilde{\pi}$. We have that for any $x \in \mathbb{R}^d$, $\|b(x) - \tilde{b}(x)\|_2 = \sigma^{-2}\|\mu - \tilde{\mu}\|_2 =: \epsilon$. Furthermore, the Hessian is $H[\log \pi] = -\sigma^{-2} I$, which implies that b is σ^{-2}-strongly concave. Therefore, per the discussion in Section 2, exponential contractivity holds with $C = 1$ and $\rho = e^{-\sigma^{-2}}$. We thus conclude that

$$\frac{C\epsilon}{\log(1/\rho)} = \frac{\sigma^{-2}\|\mu - \tilde{\mu}\|_2}{\sigma^{-2}} = \|\mu - \tilde{\mu}\|_2 = d_{\mathcal{W}}(\pi, \tilde{\pi}).$$

and hence the bound of Theorem 3.1 is tight in this setting. □

Theorem 3.1 assumes that the approximate drift is a deterministic function and that the error in the drift is uniformly bounded. We can generalize the results of Theorem 3.1 to allow for the approximate diffusion to use stochastic drift with non-uniform drift error. We will see that only the expected magnitude of the drift bias affects the final error bound. Let $\tilde{b}(\tilde{X}_t, \tilde{Y}_t)$ denote the approximate drift, which is now a function of both the current location $\tilde{X}_t$ and an independent diffusion $\tilde{Y}_t \in \mathbb{R}^\ell$:

$$\begin{aligned} \mathrm{d}\tilde{X}_t &= (\tilde{b}(\tilde{X}_t, \tilde{Y}_t))\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\tilde{W}_t^X \\ \mathrm{d}\tilde{Y}_t &= b_{aux}(\tilde{Y}_t)\,\mathrm{d}t + \Sigma\,\mathrm{d}\tilde{W}_t^Y, \end{aligned} \tag{3.2}$$

where Σ is an $\ell \times \ell$ matrix and the notation $\tilde{W}_t^X$ and $\tilde{W}_t^Y$ highlights that the Brownian motions in $\tilde{X}_t$ and $\tilde{Y}_t$ are independent. Let $\tilde{\pi}_Z$ denote the stationary distribution of $\tilde{Z}_t \triangleq (\tilde{X}_t, \tilde{Y}_t)$. For measure μ and function f, we write $\mu(f) \triangleq \int f(x)\mu(\mathrm{d}x)$ to reduce clutter. We can now state a generalization of Theorem 3.1.

Theorem 3.4 (Error induced by stochastic approximate drift)**.** *Let π and $\tilde{\pi}$ denote the invariant distributions of the diffusions in Eqs. (2.1) and (3.2), respectively. Assume that there exists a measurable function $\epsilon \in C(\mathbb{R}^d, \mathbb{R}_+)$ such that for $(\tilde{X}, \tilde{Y}) \sim \tilde{\pi}_Z$ and for all $x \in \mathbb{R}^d$,*

$$\|b(x) - \mathbb{E}[\tilde{b}(\tilde{X}, \tilde{Y}) \,|\, \tilde{X} = x]\|_2 \le \epsilon(x).$$

If the diffusion Eq. (2.1) is exponentially ergodic and the regularity conditions of Assumption 2.B hold, then

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) \le \frac{C\,\tilde{\pi}(\epsilon)}{\log(1/\rho)}.$$

Whereas the bound of Theorem 3.1 is proportional to the deterministic drift error ϵ, the bound for the diffusion with a stochastic approximate drift is proportional to the expected drift error bound $\tilde{\pi}(\epsilon)$. The bound of Theorem 3.4 thus takes into account how the drift error varies with the location of the drift. Our results match the asymptotic behavior for stochastic gradient Langevin dynamics documented in Teh et al. [38]: in the limit of the step size going to zero, they show that the stochastic gradient has no effect on the equilibrium distribution.

Example. Suppose $\tilde{Y}_t$ is an Ornstein–Uhlenbeck process with $\ell = d$, the dimensionality of $\tilde{X}_t$. That is, for some $\alpha, v > 0$, $\mathrm{d}\tilde{Y}_t = -\alpha\tilde{Y}_t\mathrm{d}t + \sqrt{2v}\,\mathrm{d}\tilde{W}_t^Y$. Then the equilibrium distribution of $\tilde{Y}_t$ is that of a Gaussian with covariance $\sigma^2 I$, where $\sigma^2 \triangleq v/\alpha$. Let $\tilde{b}(x, y) = b(x) + y$, so $\mathbb{E}[\tilde{b}(\tilde{X}, \tilde{Y}) \,|\, \tilde{X} = x] = b(x)$ and hence $d_{\mathcal{W}}(\pi, \tilde{\pi}) = 0$. □

While exponential contractivity is natural and applies in many settings, it is useful to have bounds on the Wasserstein distance of approximations when the diffusion process mixes more slowly. We can prove the analogous guarantee of Theorem 3.1 when a weaker, polynomial contractivity condition is satisfied.

Assumption 3.C (**Polynomial contractivity**)**.** *There exist constants $C > 0$, $\alpha > 1$, and $\beta > 0$ such that for all $x, x' \in \mathbb{R}^d$,*

$$d_{\mathcal{W}}(\mu_{x,t}, \mu_{x',t}) \le C\|x - y\|_2 (t + \beta)^{-\alpha}.$$

The parameters α and β determines how quickly the diffusion converges to equilibrium. Polynomial contractivity can be certified using, for example, the techniques from Butkovsky [12] (see also the references therein).

Theorem 3.5 (Error induced by approximate drift, polynomial contractivity)**.** *Let π and $\tilde{\pi}$ denote the invariant distributions of the diffusions in Eq. (2.1) and Eq. (2.2), respectively. If the diffusion Eq. (2.1) is polynomially ergodic with parameters C, α, and β, the regularity conditions of Assumption 2.B hold, and $\|b - \tilde{b}\|_2 \le \epsilon$, then*

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) \le \frac{C\epsilon}{(\alpha - 1)\beta^{\alpha - 1}}. \tag{3.3}$$

Remark 3.6 (*Coherency of the error bound*). The error bound of Eq. (3.3) has a coherent dependence on its parameters, just like Eq. (3.1). If we change the time scale of the diffusion from t to $s = at$ for some $a > 0$, the polynomial contractivity constants C, α, and β become, respectively, $C/a^\alpha, \alpha$, and β/a. Making these substitutions and replacing ϵ by ϵa, one can check that the scaling a cancels out in the error bound, so the error is independent of how we set the time scale. □

4 Overview of analysis techniques

We use Stein's method [3, 35, 37] to bound the Wasserstein distance between π and $\tilde{\pi}$ as a function of a bound on $\|b - \tilde{b}\|_2$ and the mixing time of π. We describe the analysis ideas for the setting when $\|b - \tilde{b}\|_2 < \epsilon$ (Theorem 3.1); the analysis with stochastic drift (Theorem 3.4) or assuming polynomial contractivity (Theorem 3.5) is similar. All of the details are in the Supplementary Material.

For a diffusion $(X_t)_{t\geq 0}$ with drift b, the corresponding infinitesimal generator satisfies

$$\mathcal{A}_b\phi(x) = b(x)\cdot\nabla\phi(x) + \Delta\phi(x)$$

for any function ϕ that is twice continuously differentiable and vanishing at infinity. See, e.g., Ethier and Kurtz [19] for an introduction to infinitesimal generators. Under quite general conditions, the invariant measure π and the generator $\mathcal{A}_b$ satisfy

$$\pi(\mathcal{A}_b\phi) = 0.$$

For any measure ν on $\mathbb{R}^d$ and set of test functions $\mathcal{F} \subseteq C^2(\mathbb{R}^d, \mathbb{R})$, we can define the *Stein discrepancy* as:

$$\mathcal{S}(\nu, \mathcal{A}_b, \mathcal{F}) \triangleq \sup_{\phi\in\mathcal{F}} |\pi(\mathcal{A}_b\phi) - \nu(\mathcal{A}_b\phi)| = \sup_{\phi\in\mathcal{F}} |\nu(\mathcal{A}_b\phi)|.$$

The Stein discrepancy quantifies the difference between ν and π in terms of the maximum difference in the expected value of a function (belonging to the transformed test class $\{\mathcal{A}_b\phi \,|\, \phi\in\mathcal{F}\}$) under these two distributions. We can analyze the Stein discrepancy between π and $\tilde{\pi}$ as follows. Consider a test set $\mathcal{F}$ such that $\|\nabla\phi\|_2 \leq 1$ for all $\phi\in\mathcal{F}$, which is equivalent to having $\|\phi\|_L \leq 1$. We have that

$$\begin{aligned}
\mathcal{S}(\tilde{\pi}, \mathcal{A}_b, \mathcal{F}) &= \sup_{\phi\in\mathcal{F}} |\tilde{\pi}(\mathcal{A}_b\phi)| = \sup_{\phi\in\mathcal{F}} |\tilde{\pi}(\mathcal{A}_b\phi - \mathcal{A}_{\tilde{b}}\phi)| \\
&= \sup_{\phi\in\mathcal{F}} |\tilde{\pi}(\nabla\phi\cdot b - \nabla\phi\cdot\tilde{b})| \\
&\leq \sup_{\phi\in\mathcal{F}} |\tilde{\pi}(\|\nabla\phi\|_2\|b-\tilde{b}\|_2)| \leq \epsilon,
\end{aligned}$$

where we have used the definition of Stein discrepancy, that $\tilde{\pi}(\mathcal{A}_{\tilde{b}}\phi) = 0$, the definition of the generator, the Cauchy-Schwartz inequality, that $\|\nabla\phi\|_2 \leq 1$, and the assumption $\|b-\tilde{b}\|_2 \leq \epsilon$. It remains to show that the Wasserstein distance satisfies $d_{\mathcal{W}}(\pi,\tilde{\pi}) \leq C_\pi \mathcal{S}(\tilde{\pi}, \mathcal{A}_b, \mathcal{F})$ for some constant C_π that may depend on π. This would then allow us to conclude that $d_{\mathcal{W}}(\pi,\tilde{\pi}) \leq C_\pi\epsilon$. To obtain C_π, for each 1-Lipschitz function h, we construct the solution u_h to the differential equation

$$h - \pi(h) = \mathcal{A}_g u \tag{4.1}$$

and show that $\|\nabla u_h\|_2 \leq C_\pi \|\nabla h\|_2$.

5 Application: computational–statistical trade-offs

As an application of our results we analyze the behavior of the *unadjusted Langevin Monte Carlo algorithm* (ULA) [34] when approximate gradients of the log-likelihood are used. ULA uses a discretization of the continuous-time Langevin diffusion to approximately sample from the invariant distribution of the diffusion. We prove conditions under which we can obtain more accurate samples by using an approximate drift derived from a Taylor expansion of the exact drift.

For the diffusion $(X_t)_{t\geq 0}$ driven by drift b as defined in Eq. (2.1) and a non-increasing sequence of step sizes $(\gamma_i)_{i\geq 1}$, the associated ULA Markov chain is

$$X'_{i+1} = X'_i + \gamma_{i+1}\, b(X'_i) + \sqrt{2\gamma_{i+1}}\,\xi_{i+1}, \tag{5.1}$$

where $\xi_{i+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Recently, substantial progress has been made in understanding the approximation accuracy of ULA [11, 15, 17]. These analyses show, as a function of the discretization step size γ_i, how quickly the distribution of X'_i converges to the desired target distribution.

In many big data settings, however, computing $b(X'_i)$ exactly at every step is computationally expensive. Given a fixed computational budget, one option is to compute $b(X'_i)$ precisely and run the discretized diffusion for a small number of steps to generate samples. Alternatively, we could replace $b(X'_i)$ with an approximate drift $\tilde{b}(X'_i)$ which is cheaper to compute and run the discretized approximate diffusion for a larger number of steps to generate samples. While approximating the drift can introduce error, running for more steps can compensate by sampling from a better mixed chain. Thus, our objective is to compare the ULA chain using an exact drift initialized at some point $x^\star \in \mathbb{R}^d$ to a ULA chain using an approximate drift initialized at the same point. We denote the exact and approximate drift chains by $X'_{x^\star,i}$ and $\tilde{X}'_{x^\star,i}$, respectively, and denote laws of these chains by $\mu_i^\star$ and $\tilde{\mu}_i^\star$.

For concreteness, we analyze generalized linear models with unnormalized log-densities of the form

$$\mathcal{L}(x) \triangleq \log\pi_0(x) + \sum_{i=1}^N \phi_i(x\cdot y_i),$$

where $y_1,\ldots,y_N \in \mathbb{R}^d$ is the data and x is the parameter. In this setting the drift is $b(x) = \nabla\mathcal{L}(x)$. We take $x^\star = \arg\max_x \mathcal{L}(x)$ and approximate the drift with a Taylor expansion around $x^\star$:

$$\begin{aligned}
\tilde{b}(x) \triangleq\; & (H\log\pi_0)(x^\star)(x-x^\star) \\
& + \sum_{i=1}^N \phi_i''(x^\star\cdot y_i) y_i y_i^\top (x-x^\star),
\end{aligned} \tag{5.2}$$

where H is the Hessian operator. The quadratic approximation of Eq. (5.2) basically corresponds to taking a Laplace approximation of the log-likelihood. In

practice, higher-order Taylor truncation or other approximations can be used, and our analysis can be extended to quantify the trade-offs in those cases as well. Here we focus on the second-order approximation as a simple illustration of the computational–statistical trade-off.

In order for the Taylor approximation to be well-behaved, we require the prior π_0 and link functions ϕ_i to satisfying some regularity conditions, which are usually easy to check in practice.

Assumption 5.D (Concavity and Smoothness)**.**

1. *The function* $\log \pi_0 \in C^3(\mathbb{R}^d, \mathbb{R})$ *is strongly concave,* $\|\nabla \log \pi_0\|_L < \infty$*, and* $\|H[\partial_j \log \pi_0]\|_* < \infty$ *for* $j = 1, \dots, d$*.*
2. *For* $i = 1, \dots, N$*, the function* $\phi_i \in C^3(\mathbb{R}, \mathbb{R})$ *is strongly concave,* $\|\phi_i'\|_L < \infty$*, and* $\|\phi_i'''\|_\infty < \infty$*.*

We measure computational cost by the number of d-dimensional inner products performed. Running ULA with the original drift b for T steps costs TN because each step needs to compute $x \cdot y_i$ for each of the N y_i's. Running ULA with the Taylor approximation $\tilde{b}$, we need to compute $\sum_{i=1}^N \phi_i''(x^\star \cdot y_i) y_i y_i^\top$ once up front, which costs Nd, and then for each step we just multiply this d-by-d matrix with $x - x^\star$, which costs d. So the total cost of running approximate ULA for $\tilde{T}$ steps is $(\tilde{T} + N)d$.

Theorem 5.1 (Computational–statistical trade-off for ULA)**.** *Set the step size* $\gamma_i = \gamma_1 i^{-\alpha}$ *for fixed* $\alpha \in (0, 1)$ *and suppose the ULA of Eq. (5.1) is run for* $T > d$ *steps. If Assumption 5.D holds and* $\tilde{T}$ *is chosen such that the computational cost of the second-order approximate ULA using drift Eq. (5.2) equals that of the exact ULA, then* γ_1 *may be chosen such that*

$$d_{\mathcal{W}}^2(\mu_T^\star, \pi) = \tilde{O}\left(\frac{d}{TN}\right)$$

and

$$d_{\mathcal{W}}^2(\tilde{\mu}_{\tilde{T}}^\star, \pi) = \tilde{O}\left(\frac{d^2}{N^2 T} + \frac{d^3}{N^2}\right).$$

The ULA procedure of Eq. (5.1) has Wasserstein error decreasing like $1/N$ for data size N. Because approximate ULA can be run for more steps at the same computational cost, its error decreases as $1/N^2$. Thus, for large N and fixed T and d, approximate ULA with drift $\tilde{b}$ achieves more accurate sampling than ULA with b. A conceptual benefit of our results is that we can cleanly decompose the final error into the discretization error and the equilibrium bias due to approximate drift. Our theorems in Section 3 quantifies the equilibrium bias, and we can apply existing techniques to bound the discretization error.

6 Extension: piecewise deterministic Markov processes

We next demonstrate the generality of our techniques by providing a perturbation analysis of piecewise deterministic Markov processes (PDMPs), which are continuous-time processes that are deterministic except at random jump times. Originating with the work of Davis [16], there is now a rich literature on the ergodic and convergence properties of PDMPs [2, 6, 14, 20, 29]. They have been used to model a range of phenomena including communication networks, neuronal activity, and biologic population models (see [2] and references therein). Recently, PDMPs have also been used to design novel MCMC inference schemes. zig-zag processes (ZZPs) [7–9] are a class of PDMPs that are particularly promising for inference. ZZPs can be simulated exactly (making Metropolis-Hastings corrections unnecessary) and are non-reversible, which can potentially lead to more efficient sampling [28, 30].

Our techniques can be readily applied to analyze the accuracy of approximate PDMPs. For concreteness we demonstrate the results for ZZPs in detail and defer the general treatment of PDMPs, which includes an idealized version of Hamiltonian Monte Carlo, to the Supplementary Material. The ZZP is defined on the space $E = \mathbb{R}^d \times \mathcal{B}$, where $\mathcal{B} \triangleq \{-1, +1\}^d$. Densities on $\mathcal{B}$ are with respect to the counting measure.

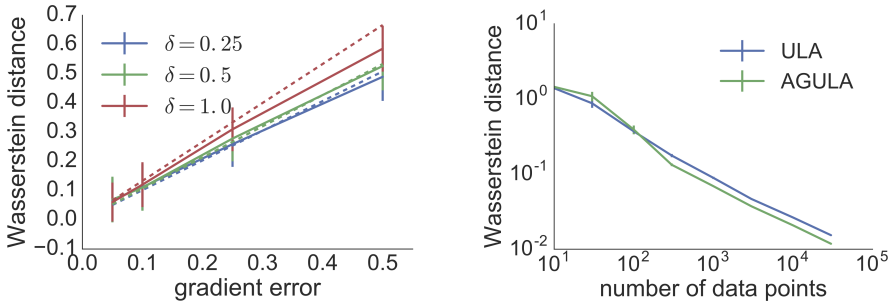
Informally, the behavior of a ZZP can be described as follows. The trajectory is X_t and its velocity is Θ_t, so $\frac{d}{dt} X_t = \Theta_t$. At random times, a single coordinate of Θ_t flips signs. In between these flips, the velocity is a constant and the trajectory is a straight line (hence the name "zig-zag"). The rate at which Θ_t flips a coordinate is time inhomogeneous. The i-th component of Θ switches at rate $\lambda_i(X_t, \Theta_t)$. By choosing the switching rates appropriately, the ZZP can be made to sample from the desired distribution. More precisely, the ZZP $(X_t, \Theta_t)_{t \geq 0}$ is determined by the switching rate $\lambda \in C^0(E, \mathbb{R}_+^d)$ and has generator

$$\mathcal{A}_\lambda \phi(x, \theta) = \theta \cdot \nabla_x \phi(x, \theta) + \lambda(x, \theta) \cdot \nabla_\theta \phi(x, \theta) \quad (6.1)$$

for any sufficiently regular $\phi : E \to \mathbb{R}$. Here $\nabla_x \phi$ denotes the gradient of ϕ with respect to x and $\nabla_\theta \phi$ is the discrete differential operator.[4] Let $(a)^+ \triangleq \max(0, a)$ denote the positive part of $a \in \mathbb{R}$ and $\partial_i \phi \triangleq \frac{\partial \phi}{\partial x_i}$. The following result shows how to construct a ZZP with invariant distribution π.

Theorem 6.1 (Bierkens et al. [9, Theorem 2.2, Proposition 2.3])**.** *Suppose* $\log \pi \in C^1(\mathbb{R}^d)$ *and* $\gamma \in$

[4] $\nabla_\theta \phi \triangleq (\partial_{\theta,1}\phi, \dots, \partial_{\theta,d}\phi)$, where $\partial_{\theta,i}\phi(x, \theta) \triangleq \phi(x, R_i\theta) - \phi(x, \theta)$ and for $i \in [d]$, the reversal function $R_i : \mathcal{B} \to \mathcal{B}$ is given by $(R_i\theta)_j \triangleq \begin{cases} -\theta_j & j = i \\ \theta_j & j \neq i. \end{cases}$



(a) (b)

Figure 1: **(a)** Gradient error ϵ versus the Wasserstein distance between π_δ and $\tilde{\pi}_{\delta,\epsilon}$, the stationary distribution of the diffusion with approximate drift $\tilde{b}_{\delta,\epsilon}(x) = \nabla \log \pi_\delta(x) + \epsilon$. The solid lines are the simulation results and the dotted lines are the theoretical upper bounds obtained from Theorem 3.1. The simulation results closely match the theoretical bounds and show linear growth in ϵ, as predicted by the theory. Due to Monte Carlo error the simulation estimates sometimes slightly exceed the theoretical bounds. **(b)** The y-axis measures the Wasserstein distance between the true posterior distribution and the finite-time distribution of the exact gradient ULA (ULA) and the approximate gradient ULA (AGULA). Except for when the number of data points $N < 100$, AGULA shows superior performance, in agreement with the analysis of Theorem 5.1. For all experiments the Wasserstein distance was estimated 10 times, each time using 1,000 samples from each distribution.

$C^0(E, \mathbb{R}^d_+)$ *satisfies* $\gamma_i(x, \theta) = \gamma_i(x, R_i\theta)$. *Let*

$$\lambda_i(x, \theta) = (-\theta_i \partial_i \log \pi(x))^+ + \gamma_i(x, \theta).$$

Then the Markov process with generator $\mathcal{A}_\lambda$ *has invariant distribution* $\pi_E(\mathrm{d}x, \theta) = 2^{-d}\pi(\mathrm{d}x)$.

Analogously to the approximate diffusion setting, we compare $(X_t, \Theta_t)_{t\geq 0}$ to an approximating ZZP $(\tilde{X}_t, \tilde{\Theta}_t)_{t\geq 0}$ with switching rate $\tilde{\lambda} \in C^0(E, \mathbb{R}^d_+)$. For example, if $\tilde{\pi}$ is an approximating density, the approximate switching rate could be chosen as

$$\tilde{\lambda}_i(x, \theta) = (-\theta_i \partial_i \log \tilde{\pi}(x))^+ + \gamma_i(x, \theta). \tag{6.2}$$

To relate the errors in the switching rates to the Wasserstein distance in the final distributions, we use the same strategy as before: apply Stein's method to the ZZP generator in Eq. (6.1). We rely on ergodicity and regularity conditions that are analogous to those for diffusions. We write $(X_{x,\theta,t}, \Theta_{x,\theta,t})$ to denote the version of the ZZP satisfying $(X_{x,\theta,0}, \Theta_{x,\theta,0}) = (x, \theta)$ and denote its law by $\mu_{x,\theta,t}$.

Assumption 6.E (ZZP polynomial ergodicity). *There exist constants* $C > 0$, $\alpha > 1$, *and* $\beta > 0$ *such that for all* $x \in \mathbb{R}^d$, $\theta \in \mathcal{B}$, *and* $i \in [d]$,

$$d_{\mathcal{W}}(\mu_{x,\theta,t}, \mu_{x,R_i\theta,t}) \leq C(t + \beta)^{-\alpha}.$$

The ZZP polynomial ergodicity condition is looser than that used for diffusions. Indeed, we only need a quantitative bound on the ergodicity constant when the chains are started with the same x value. Together with the fact that $\mathcal{B}$ is compact, this simplifies verification of the condition, which can be done using well-developed coupling techniques from the PDMP literature [2, 6, 20, 29] as well as more general Lyapunov function-based approaches [23].

Our main result of this section bounds the error in the invariant distributions due to errors in the ZZP switching rates. It is more natural to measure the error between λ and $\tilde{\lambda}$ in terms of the ℓ^1 norm.

Theorem 6.2 (ZZP error induced by approximate switching rate)**.** *Assume the ZZP with switching rate* λ *(respectively* $\tilde{\lambda}$*) has invariant distribution* π *(resp.* $\tilde{\pi}$*). Also assume that* $\int_E x^2 \pi(\mathrm{d}x, \mathrm{d}\theta) < \infty$ *and if a function* $\phi \in C(E, \mathbb{R})$ *is* π*-integrable then it is* $\tilde{\pi}$*-integrable. If the ZZP with switching rate* λ *is polynomially ergodic with constants* C, α, *and* β *and* $\|\lambda - \tilde{\lambda}\|_1 \leq \epsilon$, *then*

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) \leq \frac{C\epsilon}{(\alpha - 1)\beta^{\alpha - 1}}.$$

Remark 6.3. If the approximate switching rate takes the form of Eq. (6.2), then $\|\nabla \log \pi - \nabla \log \tilde{\pi}\|_1 \leq \epsilon$ implies $\|\lambda - \tilde{\lambda}\|_1 \leq \epsilon$. □

7 Experiments

We used numerical experiments to investigate whether our bounds capture the true behavior of approximate diffusions and their discretizations.

Approximate Diffusions. For our theoretical results to be a useful guide in practice, we would like the Wasserstein bounds to be reasonably tight and have the correct scaling in the problem parameters (e.g., in $\|b - \tilde{b}\|_2$). To test our main result concerning the error induced from using an approximate drift (Theorem 3.1), we consider mixtures of two Gaussian

densities of the form

$$\pi_\delta(x) = \frac{1}{2(2\pi)^{d/2}} \left(e^{-\|x-\delta/2\|_2^2/2} + e^{-\|x+\delta/2\|_2^2/2} \right),$$

where $\delta \in \mathbb{R}^d$ parameterizes the difference between the means of the Gaussians. If $\|\delta\|_2 < 2$, then π_δ is $(1 - \|\delta\|_2/4)$-strongly log-concave; if $\|\delta\|_2 = 2$, then π_δ is log-concave; and if $\|\delta\|_2 > 2$, then π_δ is not log-concave, but is log-concave in the tails. Thus, for all choices of δ, the diffusion with drift $b_\delta(x) \triangleq \nabla \log \pi_\delta(x)$ is exponentially ergodic. Importantly, this class of Gaussian mixtures allows us to investigate a range of practical regimes, from strongly unimodal to highly multi-modal distributions. For $d = 1$ and a variety of choices of δ, we generated 1,000 samples from the target distribution π_δ (which is the stationary distribution of a diffusion with drift $b_\delta(x)$) and from $\tilde{\pi}_{\delta,\epsilon}$ (which is the stationary distribution of the approximate diffusion with drift $\tilde{b}_{\delta,\epsilon}(x) \triangleq b_\delta(x)+\epsilon$) for $\epsilon = 0.05, 0.1, 0.25, 0.5$. We then calculated the Wasserstein distance between the empirical distribution of the target and the empirical distribution of each approximation. Fig. 1a shows the empirical Wasserstein distance (solid lines) for $\delta = 0.25, 0.5, 1.0$ along with the corresponding theoretical bounds from Theorem 3.1 (dotted lines). The two are in close agreement. We also investigated larger distances for $\delta = 1.0, 2.0, 3.0$. Here the exponential contractivity constants that can be derived from Eberle [18] are rather loose. Importantly, however, for all values of δ considered, the Wasserstein distance grows linearly in ϵ, as predicted by our theory. Results for $d > 1$ show similar linear behavior in ϵ, though we omit the plots.

Computational–statistical trade-off. We illustrate the computational–statistical trade-off of Theorem 5.1 in the case of logistic regression. This corresponds to $\phi_i(t) = \phi_{lr}(t) \triangleq -\log(1+e^{-t})$. We generate data $y_1, y_2, \ldots$ according to the following process:

$$z_i \sim \mathsf{Bern}(.5), \quad \zeta_i \sim \mathcal{N}(\mu_{z_i}, I), \quad y_i = (2z_i - 1)\zeta_i,$$

where $\mu_0 = (0, 0, 1, 1)$ and $\mu_1 = (1, 1, 0, 0)$. We restrict the domain $\mathcal{X}$ to a ball of radius 3, $\mathcal{X} = \{x \in \mathbb{R}^4 \,|\, \|x\|_2 \leq 3\}$, and add a projection step to the ULA algorithm [11], replacing Z_i' with $\arg\min_{z\in\mathcal{X}} \|Z_i' - z\|_2$. While Theorem 5.1 assumes $\mathcal{X} = \mathbb{R}^4$, the numerical results here on the bounded domain still illustrate our key point: for the same computational budget, computing fast approximate gradients and running the ULA chain for longer can produce a better sampler. Fig. 1b shows that except for very small N, the approximate gradient ULA (AGULA), which uses the approximation in Eq. (5.2), produces better performance than exact gradient ULA (ULA) with the same budget. For each data-set size (N), the true posterior distribution was estimated by running an adaptive Metropolis-Hastings (MH) sampler for 100,000 iterations. ULA and AGULA were each run 1,000 times to empirically estimate the approximate posteriors. We then calculated the Wasserstein distance between the ULA and AGULA empirical distributions and the empirical distribution obtained from the MH sampler.

8 Discussion

Related Work. Recent theoretical work on scalable MCMC algorithms has yielded numerous insights into the regimes in which such methods produce computational gains [1, 24, 25, 32, 36]. Many of these works focused on approximate Metropolis-Hastings algorithms, rather than gradient-based MCMC. Moreover, the results in these papers are for discrete chains, whereas our results also apply to continuous diffusions as well as other continuous-time Markov processes such as the zig-zag process. Perhaps the closest to our work is that of Rudolf and Schweizer [36] and Gorham et al. [22]. The former studies general perturbations of Markov chains and includes an application to stochastic Langevin dynamics. They also rely on a Wasserstein contraction condition, like our Assumption 2.A, in conjunction with a Lyapunov condition on the perturbed chain. However, our more specialized analysis is particularly transparent and leads to tighter bounds in terms of the contraction constant ρ: the bound of Rudolf and Schweizer [36] is proportional to $(1-\rho)^{-1}$ whereas our bound is proportional to $-(\log \rho)^{-1}$. Another advantage of our approach is that our results are more straightforward to apply since we do not need to directly analyze the Lyapunov potential and the perturbation ratios as in Rudolf and Schweizer [36]. Our techniques also apply to the weaker polynomial contraction setting. Gorham et al. [22] have results of similar flavor to ours and also rely on Stein's method, but their assumptions and target use cases differ from ours. Our results in Section 5, which apply when ULA is used with a deterministic approximation to the drift, complement the work of Teh et al. [38] and Vollmer et al. [39], which provides (non-)asymptotic analysis when the drift is approximated stochastically at each iteration.

Conclusion. We have established general results on the accuracy of diffusions with approximate drifts. As an application, we show how this framework can quantify the computational–statistical trade-off in approximate gradient ULA. The example in Section 7 illustrates how the log-concavity constant can be estimated in practice and how theory provides reasonably precise error bounds. We expect our general framework to have many further applications. In particular, an interesting direction is to extend our framework to analyze the trade-offs in subsampling Hamiltonian Monte Carlo algorithms and stochastic Langevin dynamics.

Acknowledgments

Thanks to Natesh Pillai for helpful discussions and to Trevor Campbell for feedback on an earlier draft. JHH was partially supported by the U.S. Government under FA9550-11-C-0028 and awarded by the DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

References

[1] P. Alquier, N. Friel, R. Everitt, and A. Boland. Noisy Monte Carlo: convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 26:29–47, 2016.

[2] R. Azaïs, J.-B. Bardet, A. Génadot, N. Krell, and P.-A. Zitt. Piecewise deterministic Markov process — recent results. *ESAIM: Proceedings*, 44: 276–290, Jan. 2014.

[3] A. D. Barbour. Stein's Method for Diffusion Approximations. *Probability theory and related fields*, 84:297–322, 1990.

[4] R. Bardenet, A. Doucet, and C. C. Holmes. Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *International Conference on Machine Learning*, pages 405–413, 2014.

[5] R. Bardenet, A. Doucet, and C. C. Holmes. On Markov chain Monte Carlo methods for tall data. *arXiv.org*, May 2015.

[6] M. Benaïm, S. Le Borgne, F. Malrieu, and P.-A. Zitt. Quantitative ergodicity for some switched dynamical systems. *Electronic Communications in Probability*, 17(0), 2012.

[7] J. Bierkens and A. Duncan. Limit theorems for the Zig-Zag process. *arXiv.org*, July 2016.

[8] J. Bierkens and G. O. Roberts. *A piecewise deterministic scaling limit of Lifted Metropolis-Hastings in the Curie-Weiss model*. *The Annals of Applied Probability*, 2016.

[9] J. Bierkens, P. Fearnhead, and G. O. Roberts. The Zig-Zag Process and Super-Efficient Sampling for Bayesian Analysis of Big Data. *arXiv.org*, July 2016.

[10] F. Bolley, I. Gentil, and A. Guillin. Convergence to equilibrium in Wasserstein distance for Fokker–Planck equations. *Journal of Functional Analysis*, 263(8):2430–2457, Oct. 2012.

[11] S. Bubeck, R. Eldan, and J. Lehec. Finite-Time Analysis of Projected Langevin Monte Carlo. In *Advances in Neural Information Processing Systems*, July 2015.

[12] O. Butkovsky. Subgeometric rates of convergence of Markov processes in the Wasserstein metric. *The Annals of Applied Probability*, 24(2):526–552, Apr. 2014.

[13] T. Chen, E. B. Fox, and C. Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, 2014.

[14] O. L. V. Costa and F. Dufour. Stability and Ergodicity of Piecewise Deterministic Markov Processes. *SIAM Journal on Control and Optimization*, 47(2):1053–1077, Jan. 2008.

[15] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017.

[16] M. H. A. Davis. Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1984.

[17] A. Durmus and E. Moulines. Sampling from a strongly log-concave distribution with the Unadjusted Langevin Algorithm. *HAL*, pages 1–25, Apr. 2016.

[18] A. Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, pages 1–36, Oct. 2015.

[19] S. N. Ethier and T. G. Kurtz. Markov processes: characterization and convergence, volume 282. John Wiley & Sons, 2009.

[20] J. Fontbona, H. Guérin, and F. Malrieu. Quantitative estimates for the long-time behavior of an ergodic variant of the telegraph process. *Advances in Applied Probability*, 44:977–994, 2012.

[21] R. Ge and J. Zou. Rich Component Analysis. *arXiv.org*, July 2015.

[22] J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring Sample Quality with Diffusions. *arXiv.org*, Nov. 2016.

[23] M. Hairer, J. C. Mattingly, and M. Scheutzow. Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations. *Probability theory and related fields*, 149(1-2):223–259, Oct. 2009.

[24] J. E. Johndrow, J. C. Mattingly, S. Mukherjee, and D. Dunson. Approximations of Markov Chains and High-Dimensional Bayesian Inference. *arXiv.org*, Aug. 2015.

[25] J. E. Johndrow, J. C. Mattingly, S. Mukherjee, and D. Dunson. Approximations of Markov Chains and Bayesian Inference. *arXiv.org*, stat.CO:1–53, Jan. 2016.

[26] A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget. In *International Conference on Machine Learning*, 2014.

[27] D. Maclaurin and R. P. Adams. Firefly Monte Carlo: Exact MCMC with Subsets of Data. In *Uncertainty in Artificial Intelligence*, Mar. 2014.

[28] A. Mira and C. J. Geyer. On non-reversible Markov chains. *Monte Carlo Methods, Fields Institute/AMS*, pages 95–110, 2000.

[29] P. Monmarché. On $\mathcal{H}^1$ and entropic convergence for contractive PDMP. *Electronic Journal of Probability*, 20:1–30, 2015.

[30] R. M. Neal. Improving asymptotic variance of MCMC estimators: Non-reversible chains are better. Technical Report 0406, University of Toronto, 2004.

[31] R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman and Hall/CRC, 2011.

[32] N. S. Pillai and A. Smith. Ergodicity of Approximate MCMC Chains with Applications to Large Data Sets. *arXiv.org*, May 2014.

[33] M. Quiroz, M. Villani, and R. Kohn. Scalable MCMC for Large Data Problems using Data Subsampling and the Difference Estimator. *arXiv.org*, July 2015.

[34] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, Nov. 1996.

[35] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

[36] D. Rudolf and N. Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *arXiv.org*, Mar. 2015.

[37] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602, 1972.

[38] Y. W. Teh, A. H. Thiery, and S. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17(7):1–33, Mar. 2016.

[39] S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. (Non-) asymptotic properties of Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research*, 17(159):1–48, 2016.

[40] M. Welling and Y. W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *International Conference on Machine Learning*, 2011.